Data mining for qualitative dataset using association rules

A Synopsis

Submitted in

partial fulfillment for the award of the degree

of

Doctor of philosophy

(Computer Science Engineering)

Supervised by

Dr. Amita Sharma

Submitted by

JV'n Mr Amit Kumar Chandanan

Department of Computer Science and Information Technology
Faculty of Engineering & Technology
Jayoti Vidyapeeth Women's University
Jaipur (Rajasthan), India
March 2012

Introduction

Currently the world has a wealth of data, stored all over the planet, but we need to understand that data. It has been stated that the amount of data doubles approximately every twenty months [1]. This is especially true since the use of computers and electronic database packages. The amount or quantity of data easily exceeds what a human can comprehend on their own and thus if we wish to use and understand as much data as possible we need tools to help us. From this overwhelming state, the field of data mining has taken off and become hotly utilized.

Association rule mining is one of the dominating data mining technologies. Association rule mining is a process for finding associations or relations between data items or attributes in large datasets. It allows popular patterns and associations, correlations, or relationships among patterns to be found with minimal human effort, bringing important information to the surface for use. Association rule mining has been proven to be a successful technique for extracting useful information from large datasets. Various algorithms or models were developed many of which have been applied in various application domains that include telecommunication networks, market analysis, risk management, inventory control and many others. The success of applying the extracted rules to solving real world problems is very often restricted by the quality of the rules. However, the quality of the extracted rules has not drawn adequate attention [3] [8]. Measuring the quality of association rules is also difficult and current methods appear to be unsuitable, especially when multi-level and cross level rules are involved. Moreover, most of the successful applications are restricted to cases where the datasets involve only a single concept level and the success of the application is heavily dependent on the quality of the discovered rule set. Mining quality non-redundant multi-level association rules from multi-level datasets is a challenge still needing to be worked on and it is a desired goal for helping to solve real world problems.

Review of literature

Data mining and the use of the discovered data and knowledge is a major field of research. Research in this field is also often taken and applied to real world scenarios. In order to improve the usage of data and the knowledge it contains, it is necessary to develop better techniques. Improvements to rule mining, being one of the major data mining technologies, would result in benefits to many applications. Hence the development of new novel techniques to discover and mine high quality association rules from multi-level datasets and effectively use them is important. Furthermore, to ensure that high quality rules can be identified, it is equally as important to have an assessment for measuring a rule's usefulness or interestingness. By achieving this, then the applications that utilize association rules can be improved.

For most of the work done in developing association rule mining, the primary focus has been on the efficiency of the approach (how quickly can it derive the rules) and to a lesser extent the quality (and the evaluation process / measures to determine the quality) of the derived rules has been emphasized. Often for a dataset, a huge number of rules can be derived, but many of them (in some cases a significant number) can be redundant to other rules and thus are useless in practice. The extremely large number of rules makes it difficult for the end users to comprehend and therefore effectively use the discovered rules and thus significantly reduces the effectiveness of rule mining algorithms. If the extracted knowledge can't be effectively used in solving real world problems, the effort of extracting the knowledge is worth little. This is a serious problem but not yet solved satisfactorily. Some approaches aiming at reducing the number of extracted rules and also eliminating redundancy have been proposed [8]. However, none of them deal with multi-level datasets and instead have focused solely on datasets with only a single level. The traditional approaches do not take into account that there may be a data hierarchy so they find associations at just one level of the hierarchy and fail to discover associations and rules for data in other hierarchical levels or even rules that span across two or more levels. For such multi-level datasets, the issue of deriving rules is even more serious, since the presence of multiple concept levels in a dataset means that items and therefore patterns will be at multiple levels, which inevitably increases the number of rules that can be discovered.

The problems of finding frequent item sets are basic in multi level association rule mining, fast algorithms for solving problems are needed. This paper presents an efficient version of Apriori algorithm for mining multi-level association rules in large databases to finding maximum frequent itemset at lower level of abstraction [9]. These methods reduce the time and increase the throughput, but the redundancy in the dataset left for the work.

Association rule mining aims to extract interesting correlations and associations among sets of items in large datasets. The extracted knowledge has been proved useful in solving real world problems. However, the overwhelming number of extracted association rules and the redundancy existing in the extracted rules are two obstacles that seriously affect the effective use of the extracted knowledge in solving real world problems.

In fact, there is a diversity of different measures, which include topics such as conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility and actionability. These topics are the most commonly used to evaluate the interestingness of association rules [4]. Most of this work in measuring a rule's interestingness has come about because using the support and confidence values of a rule were insufficient [5]. Thus a better measure for determining the quality of an

association rule (or rule set) is needed, especially for multi-level and cross level rules as they differ to single level rules and most existing work has focused on single level rule sets.

The first problem undertaken by this research program is to develop a measure to evaluate the interestingness of discovered association rules, especially multi-level and cross level rules, which takes into account the underlying hierarchy / taxonomy. Most work has focused on association rules discovered from single level datasets.

The Web is growing rapidly. More and more information is being made available electronically. Users are often overwhelmed by the huge amount of information and are faced with the great challenge of finding the most relevant pieces of information in a short amount of time. Significant research endeavors are being invested into building support tools that ensure the right information is delivered to the right people at the right time.

The second problem undertaken by this research program aims to extend recommender systems to working with association rules extracted from the domain taxonomy data that are not only the traditional single-level rule sets but also multi-level and cross-level rule sets. This proposed approach attempts to provide:

1. An automated approach to discover non-redundant association rules from multi-level datasets.
2. Techniques, which consider the underlying hierarchy / taxonomy of the dataset, to measure the interestingness of rules derived from a multi-level dataset.

In short, the proposed work focuses on obtaining a set of non-redundant rules, determining their interestingness or quality and effectively applying them into a recommender system [7].

Justification /Importance of study

The aim of this thesis was to make a contribution to the field of association rule mining, recommender systems and the general field of data mining in the areas of using multi-level datasets, the evaluation of the quality of association rules and the use of association rules in recommender systems. This study has investigated and developed an effective method for mining non-redundant association rules, alternative approaches to measure the interestingness of a multi-level or cross-level association rule and/or rule set and an approach for utilizing them effectively in a recommender system, in order to solve the cold-start problem. The primary contributions for this research are:

- Comprehensive review and survey of current association rule mining techniques along with a proposal for a novel approach to remove hierarchically redundant association rules without information loss from rules discovered from a multi-level dataset.
- Comprehensive review and survey of measures used to evaluate the interestingness and/or quality of association rules and rule sets, along with a proposal for new measures that consider the hierarchy / taxonomy structure of the dataset, to evaluate the interestingness of association rules discovered from a multi-level dataset.
- New diversity based measures allowing the difference between items within a multi-level or cross-level association rule to be determined. These measures allow either the overall diversity of a rule to be determined or the diversity between a rule's antecedent and consequent to be determined.
- A peculiarity based distance measure that measures the distance between two association rules from a multi-level dataset which takes into account any taxonomic relationships between items in the rules.

Research objectives and Problem

A single level dataset (sometimes known as a flat dataset) is one where there is no taxonomy or concept hierarchy involved in the dataset. All of the items objects or attributes in such a dataset are unrelated; none are a super-topic or sub-topic of another and the attributes are generally accepted to be independent of each other. As mentioned it is often possible to extract a large number of association rules from a dataset, but many of them can be redundant. In this section we look at determining and removing these redundant association rules. Firstly we define the problem, then define redundancy, following that we detail concise representations and finally the algorithms for non-redundant rule extraction in single level datasets.

Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of unique items with m members, t is a transaction that contains a set of items from I so that $t \subseteq I$ and T is a database or dataset containing a set of identifiable transactions t. As mentioned an association rule is implication of the following form $X \Rightarrow Y$, where $X, Y \subset I$ and are known as itemsets, containing a set of items where $X \cap Y = \phi$. The closure operation of the Galois connection provides the definition for closed itemsets [6]. The definition is as follows; $\forall i \in I$ and $\forall t \in T$, if an item i appear in a transaction t, then both i and t have a binary relation denoted by $i\delta t$. The following mappings define the Galois connection of the binary relation, where $X \subseteq I$, $Y \subseteq T$:

$$\tau: 2^I \rightarrow 2^T, \tau(X) = \{ t \in T | \forall i \in X, i\delta t \}$$

$$\tau: 2^T \rightarrow 2^I, \tau(X) = \{ i \in I | \forall t \in Y, i\delta t \}$$

From this $\tau(X)$ is known as the transaction mapping of *X*, while $\gamma(Y)$ is known as the item aping of *Y*. Then the closure of *X* is $\gamma \circ \tau(X)$, giving the common items among the transactions, each of which contain *X*

Usually there are two subproblems that association rule mining can be decomposed into; finding frequent itemsets which have a support greater than or equal to a predefined minimum support and using those frequent itemsets to generate association rules which satisfy both the minimum support and minimum confidence thresholds.

Methodology

With the objectives or goals and aims of the research stated it is necessary to design a strategy and methodology to achieve these goals. A research design can be described as being a plan for getting from one point to another in which a logical sequence of steps or tasks that details the path that connects an initial research question or problem to its conclusion. The steps involved in the research process are as follow:

- The literature review and low level feature extraction and study existing methods
- To find the research problem from the study
- Design algorithm and theory and develop a system that support for experiments
- Data collection and experiment with these data and evaluate or analyze the result of experiment
- Derive Conclusion from the experimental result and discover a new knowledge from the results

Place of work and facilities available

Place of Work: Jabalpur M P & Jayoti Vidyapeeth Women's University, Jaipur
Facility required: Computer with Matlab and Weka data mining tool

References

[1] Goethals, O. Maimon & L. Rokach , "*The Data Mining and Knowledge Discovery Handbook"*, Springer Science Business Media ,New York, pp. 377-397 , 2005

[2] Pasquier, N., Taouil, R., Bastide, Y., Stumme, G., & Lakhal, L. ,"*Generating a Condensed Representation for Association Rules*", Journal of Intelligent Information Systems, *24*(1), 29-60, 2005.

[3] Xu, Y., Li, Y, "*Mining for Useful Association Rules Using the ATMS"*, Paper presented at the International Conference on Computational Intelligence for Modelling, Control and Automation, Vienna, Austria, 28-30 Nov, 2006.

[4] Geng, L., & Hamilton, H. J., *"Interestingness Measures for Data Mining: A Survey"*. ACM Computing Surveys (C SUR), *38*(3), 2006.

[5] McGarry, K. ,*"A Survey of Interestingness Measures for Knowledge Discovery"*, The Knowledge Engineering Review, *20*(1), 39-61, 2006.

[6] Ganter, B., & Wille, R, *"Formal Concept Analysis: Mathematical Foundations"*, Springer-Verlag, New York, 2005.

[7] Aciar, S., Zhang, D., Simoff, S., & Debenham, J. "*Recommender System Based on Consumer Product Reviews*", Paper presented at the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 2006.

[8] Xu, Y., & Li, Y. ,*"Generating Concise Association Rules"*, Paper presented at the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM'07), Lisbon, Portugal, 6-8, Nov, 2007.

[9] Gautam P ,Pardasi K R," *Algorithm for efficient multilevel association mining* ", International Journal on Computer Science and Engineering, Vol. 02,No 05, 2010 ,1700-1704.